
HANS'2000 und Zeichensätze

Antrag an die HANS-Nutzergemeinschaft

Thomas Berger

<ThB@gymel.com>

27.1.2003

Inhaltsverzeichnis

1. Antrag	.. 1
1.1. Vorschlag	.. 1
1.2. Konsequenzen	.. 1
1.3. Alternativen	.. 2
2. Begründung	.. 2
2.1. Geschichte	.. 2
2.2. Entscheidungskriterien	.. 3
3. Handreichungen	.. 3
Quellen	.. 3

1. Antrag

1.1. Vorschlag

HANS'2000 wird den OstWest-Zeichensatz von allegro-C benutzen, wie er in der „Amtlichen Tabelle der Standard-Zeichencodes (sog. OstWest-Zeichensatz)“ [<http://www.allegro-c.de/zcodes.htm>] definiert ist.

Dieser Zeichensatz realisiert im Wesentlichen das in *DIN 31628* als Stufe 2 definierte Repertoire.

1.2. Konsequenzen

1. Gewisse Zeichen werden nun korrekt indexieren
2. Ein Datenaustausch mit bibliothekarischen Anwendungen kann auf Zeichenebene relativ verlustfrei realisiert werden.
3. Die Installation der True-Type-Fonts *allegro Arial* (*a-arial.ttf*), *ALLEGROLG* (allegro Letter Gothic, *a-letter.ttf*) und *allegro New Roman* (*a-times.ttf*) ist für den Betrieb der Windows-Module a99 und alcarta notwendig.

Die Installation des Systemfonts *a-dos.fon* und / oder Nutzung des DOS-Fonts *OSTWEST.FON* (mittels des Hilfsprogramms *fontload.com* zu laden) wird für die Nutzung der DOS-Module von allegro-C empfohlen.

4. Da die folgenden Diakritika im OstWest-Zeichensatz isoliert vorliegen und als Prototypen für die Montage kombinierter Zeichen in Exporten dienen, sind insgesamt 576 Zeichen darstellbar. Die isolierten Diakritika sind: Akut, Gravis, Zirkumflex, übergesetzter Punkt, Trema/Diärese, Brevis, Doppelakut, Hacek, übergesetzter Ringel, Makron, Brevis untergesetzt, Ogonek, Cedille, Punkt untergesetzt, Schrägstrich durchgezogen, Querstrich durchgezogen, Unterstreichung (Makron untergesetzt), Tilde.

5. Für das Paragraphenzeichen § gab es bislang zwei Eingabemöglichkeiten, die offizielle mit Code 21 sowie die durch die PC-Tastatur normalerweise gelieferte Zeichen mit Code 245. Letzteres ist nicht mehr zulässig, denn Code 245 ist belegt durch ŋ (kleines n mit Akut). Vorhandene Daten müssen durch die Anwender per globaler Ersetzung in den betroffenen Datensätzen umgesetzt werden.

Über die Tastatur eingebare akzentuierte Großbuchstaben (etwa Â) liefern weiterhin falsche Ergebnisse, ansonsten kann die Tastatur weiterhin „normal“ benutzt werden.

6. Das Zeichen 249, dargestellt entweder als fetter mittiger Punkt • oder als Trema ˆ ist belegt durch das Zeichen ś (kleines s mit Akut). Dieses Zeichen ist lt. HANS-Datenformat [1] [3] ein Steuerzeichen in den HANS-Kategorien #100ff bzw. #120ff sowie #713ff und #714ff, das dazu dient, eine „Vorlageform“ einer verknüpften Ansetzungsform festzuhalten. Nach aktuellem Kenntnisstand gibt es nur einen Datenbestand, in dem diese Erfassungsvariante teilweise genutzt wurde.

Vorhandene Daten der Anwender können mit der mit HANS'2000 mitgelieferten Routine `g-alt249.hpr` eine *globale Manipulation* zur einmaligen Umsetzung des betroffenen Teilbestands durchführen.

1.3. Alternativen

Verzicht auf Kompatibilität mit anderen bibliothekarischen Anwendungen und aufwendige Umarbeitung des Hilfesystems von a99 auf den Zeichensatz CP 1252

2. Begründung

2.1. Geschichte

Das HANS-Datenformat [1] [3] und seine Vorgängerdokumente *HANSEATICS Nr. 2* vom September 1995 und *HANSEACTICS* vom Winter 1992/93 machen keine Aussage zum Zeichensatz. Implizit ist jedoch die Unterstellung, dass der Anwendung der Zeichensatz CP 437 zugrunde liegt, der bis einschließlich MS-DOS 5.0™ der Standardzeichensatz für IBM PC's war.

Seit der Markteinführung von MS-DOS 6™ ca. 1992 ist der Standardzeichensatz (für DOS-basierende Anwendungen) auf PC's der Zeichensatz CP 850¹. Parallel dazu propagierte die allegro-Entwicklungsabteilung seit 1993 einen *bibliothekarischen* Zeichensatz, der seit 1994 OstWest-Font oder -Zeichensatz heisst. Dieser Zeichensatz ist kompatibel zu CP 437, nicht jedoch zu CP 850.

Windows-Rechner beherrschen über eine interne Umsetzung die Zeichenkonversion zwischen dem von Windows-Programmen benutzten Zeichensatz CP 1252 und der CP 850 der DOS-Anwendungen. Analog dazu besitzen die Windows-Module von allegro-C Konversionstabellen, um zwischen den von ihnen genutzten Nichtstandard-Windows-Codierungen („allegro-Windows“) und der OstWest- Codierung der zugrundeliegenden Datenbank zu transformieren.

Kein bekannter bibliothekarischer Zeichensatz wird von den gängigen Betriebssystemen unterstützt oder ist gängigen Konversionsprogrammen bekannt. Dies gilt für allegro-Ostwest und den PICA-Zeichensatz sowie für

¹Dies blieb in Deutschen Universitätsbibliotheken bis zum Ende der 90er Jahre aus weitgehend unbemerkt, weil in den dortigen lokalen Netzwerken erst spät auf Windows-Betriebssysteme umgestiegen wurde.

die international normierten Zeichenätze ISO 5486-1983 (für MAB2) und ANSEL (für USMARC). Unicode™ ist aus Gründen der Kompatibilität mit vorhandenen (DOS-)Datenbanken und aus Rücksicht auf Anwender mit DOS, Windows-3.1 und Windows'9x-Betriebssystem auch mittelfristig keine Option für die Allegro-Entwicklungsabteilung.

2.2. Entscheidungskriterien

Sowie Datenaustausch mit anderen Anwendungen möglich sein soll, ist es unabdingbar, einen definierten Zeichensatz zu benutzen, dies gilt im Prinzip bereits für die Anzeige auf dem Bildschirm und die Einabe der Daten (wie oben erwähnt, besteht eine offensichtliche Inkompatibilität zwischen CP 437 und CP 850 in der Codierung des Zeichens hinter der Paragraphen-Taste §).

Alternativen sind primär nur der Rückzug auf den nativen Systemzeichensatz CP 850 oder der Einsatz des allegro-offiziösen OstWest-Zeichensatzes. Erstere Entscheidung bedeutet weitgehende Interoperabilität mit Windows-Anwendungen, letztere die Kompatibilität mit bibliothekarischen Anwendungen.

HANS'2000 übernimmt vom Standardsystem von allegro-C ca. 250 Hilfetexte als RTF-Dateien, darunter sind wesentliche Teile des allegro-Systemhandbuchs. In Folge einer Abweichung vom allegro-Standard müssten diese Dateien umständlich separat gepflegt werden. Dasselbe gilt für einige Dutzend Zeichencodierungs- und Exporttabellen.

Obwohl Cut & Paste durch den Einsatz eines nicht-nativen Zeichensatzes nur eingeschränkt möglich ist, bestehen aus dem OstWest-Zeichensatz heraus weitgehende Exportmöglichkeiten auch in Windows- oder WWW-Anwendungen, die weit mehr als die ca. 220 möglichen Zeichen eines „Zeichensatzes“ abdecken. Zum OstWest-Zeichensatz äquivalente Daten liegen in den Bibliotheksverbänden und den nationalen Normdateien in großer Zahl vor und mussten bislang beim Datenimport in HANS-Datenbanken künstlich in ihrem Zeichenrepertoire reduziert werden.

3. Handreichungen

Werkzeuge

Betriebssystem Windows-Zeichentabelle im Startmenü / Programme / Zubehör

a99 / alcarta Diverse Hilfeseiten, insbes. `charger.rtf` und `spchr.rtf`.

PRESTO und andere DOS-Programme Das (beim Start der DOS-Programme von allegro-C automatisch geladene) Hilfsprogramm `aw.exe` zeigt bei Betätigen der Tastenkombination **Alt-w** eine Zeichentabelle in einem kleinen Fenster.

Zeichentabellen

OstWest-Zeichensatz <http://www.gymel.com/charsets/allegro-ostwest.html>

CP 437 <http://www.gymel.com/charsets/CP437.html>

CP 850 <http://www.gymel.com/charsets/CP850.html>

Unterschiede zwischen OstWest bzw. CP 437 einerseits und CP 850 andererseits <http://www.gymel.com/charsets/legacy.html>

Quellen

[1] Harald Weigel. Thomas Berger. *HANS : Datenformat. HANS-Datenformat. 1996-*.
Versionsgeschichte

Version [1] Stand: 26.8.1996

Version [...] 1998/99

Version [5] 19.1.2000

Letzte Überarbeitung

Online unter <ftp://ftp.sub.uni-hamburg.de/pub/hans/misc/doku/hnsdform.pdf>.